Journal of Classification 5:249-282 (1988)

A Maximum Likelihood Methodology for Clusterwise Linear Regression

Wayne S. DeSarbo
University of Michigan

William L. Cron
Southern Methodist University

Abstract: This paper presents a conditional mixture, maximum likelihood methodology for performing clusterwise linear regression. This new methodology simultaneously estimates separate regression functions and membership in K clusters or groups. A review of related procedures is discussed with an associated critique. The conditional mixture, maximum likelihood methodology is introduced together with the E-M algorithm utilized for parameter estimation. A Monte Carlo analysis is performed via a fractional factorial design to examine the performance of the procedure. Next, a marketing application is presented concerning the evaluations of trade show performance by senior marketing executives. Finally, other potential applications and directions for future research are identified.

Keywords: Cluster analysis; Multiple regression; Maximum likelihood estimation; E-M algorithm; Marketing trade shows.

1. Introduction

Ordinary least-squares (OLS), or multiple regression, has been frequently utilized in social science research to summarize the relationship

---

We wish to thank the editor and three anonymous reviewers for their insightful comments which helped to improve this manuscript.

Authors' Addresses: Wayne S. DeSarbo, Departments of Marketing and Statistics, Business School of the University of Michigan, Ann Arbor, MI 48104, USA, and William L. Cron, Department of Marketing, Edwin L. Cox School of Business, Southern Methodist University, Dallas, TX 75275, USA.

between a predesignated set of independent variables and a single dependent variable. Let:

i = 1.....I subjects or observations;
j = 1.....J independent variables;
y_i = the value of the dependent variable for subject/observation i;
X_{ij} = the value of the j-th independent variable for subject/observation i;
b_j = the j-th OLS regression coefficient;
e_i = error for subject/observation i.

Then, the standard OLS linear regression model can be expressed as:

$$y_i = \sum_{j=1}^{J} X_{ij} b_j + e_i \tag{1}$$

or

$$\mathbf{y} = \mathbf{Xb} + \mathbf{e}\,, \tag{2}$$

where $\mathbf{y} = ((y_i))$, $\mathbf{X} = ((X_{ij}))$, $\mathbf{b} = ((b_j))$, and $\mathbf{e} = ((e_i))$. Given an independent sample of subjects/observations for $\mathbf{y}$ and $\mathbf{X}$, one is typically interested in estimating b_j in order to minimize the following error sums of squares:

$$\underset{b_j}{Min}\, Z = \sum_{i=1}^{I} \left[y_i - \sum_{j=1}^{J} X_{ij} b_j \right]^2$$

$$= \sum_{i=1}^{I} e_i^2 \,. \tag{3}$$

Johnston (1984) and others have derived the well known analytical expression for estimating $\mathbf{b}$ that minimizes (3):

$$\hat{\mathbf{b}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \,. \tag{4}$$

Maddala (1976) and others that show if the assumption is made that the random vector $\mathbf{e}$ is multivariate normally distributed, then the likelihood function can be written (assuming $E(\mathbf{ee}') = \sigma^2\mathbf{I}$, where $\mathbf{I}$ is an identity matrix) as:

$$L(\mathbf{y} \mid \mathbf{b}, \sigma^2) = (2\pi\sigma^2) \exp\left[-\frac{(\mathbf{y} - \mathbf{Xb})'(\mathbf{y} - \mathbf{Xb})}{2\sigma^2} \right], \tag{5}$$

TABLE 1

Synthetic Regression Data

i	X_1	X_2	Y	
1	1	-3	-5	
2	1	-2	-3	
3	1	-1	-1	
4	1	0	1	GROUP 1
5	1	1	3	$y_i = 2X_{2i}+1$
6	1	2	5	
7	1	3	7	
8	1	-3	5	
9	1	-2	3	
10	1	-1	1	
11	1	0	-1	GROUP 2
12	1	1	-3	$y_i = -2X_{2i}-1$
13	1	2	-5	
14	1	3	-7	

and the corresponding maximum likelihood estimates for **b** that maximize the likelihood function in (5) are equivalent to those obtained from least squares estimation (i.e., in expression (4)).

There are many applications that arise in the social and physical sciences, however, where the estimation of a single set of regression coefficients may prove to be "misleading." Consider, for example, the small illustrative, synthetic data set provided in Table 1, with $J = 2$ and $I = 14$. If one were to estimate one regression function for all 14 subjects/observations, the resulting estimated linear function would be:

$$\hat{y}_i = 0X_{2i} + 0, \tag{6}$$

which naturally renders an $R^2 = 0$, a very poor summary of the structure of the data displayed in Table 1. As seen in Table 1, if one were initially to cluster the observations/subjects into two groups, where group one was comprised of observations/subjects 1-7 and group two contained observations/subjects

8-14, and estimate two separate cluster regression functions, then the functions would be:

$$\text{Group 1: } y_i = 2X_{2i} + 1 \qquad i = 1, \ldots, 7$$
$$\text{Group 2: } y_i = -2X_{2i} - 1 \qquad i = 8, \ldots, 14, \tag{7}$$

with a combined $R^2 = 1.00$ indicating a perfect fit. Thus, the single estimated regression function in (6) "misrepresents" or "masks" the true structure present in the data. While one could legitimately argue for first plotting the data prior to estimation to check for such structure, such graphical displays cannot easily detect such "clusterings" as the dimensionality of the problem (J) increases. In addition, in many types of response surface estimation applications via experimental designs involving replications within subjects (e.g., conjoint analysis studies in marketing), the independent variable set often remains constant from subject to subject making such graphical detection extremely difficult.

This paper presents a new methodology for simultaneously estimating clusters and corresponding separate cluster regression functions given **X** and **y** from a sample of independent observations/subjects. We utilize finite conditional normal mixture distributions in a maximum likelihood context to estimate these parameters. We first review existing procedures that attempt to derive such simultaneous estimates. Next, the new methodology is presented together with the technical details of the E-M algorithm utilized for estimation. A Monte Carlo analysis is presented to examine the performance of this new methodology as a number of data and program options are experimentally manipulated. A marketing application is presented to examine the different evaluative criteria various senior managers utilize to evaluate the performance of their participation in trade shows. Finally, other applications as well as directions for future research are provided.

2. Literature Review

Much of the related psychometric and classification literature concerns attempts to rescale simultaneously the input variables and to solve for some clustering, all to optimize a common objective function. For example, DeSarbo, Carroll, Clark, and Green (1984) have devised the SYNCLUS methodology which simultaneously solves for a partitioning and a set of rescaling constants for the variables, all to optimize one common objective function. DeSarbo and Mahajan (1984) generalize this SYNCLUS methodology to accommodate constraints, different types of clustering schemes, and a general linear transformation of the variables. De Soete, DeSarbo, and Carroll

(1985) have extended these concepts to an optimal variable weighting scheme for hierarchical clustering where both variable weights and ultrametric trees are simultaneously estimated. Note, however, that none of these approaches are appropriate for a clusterwise regression context with dependent and independent variables.

The term "clusterwise regression" was originally coined by Späth (1979, 1981, 1982, 1985). Späth developed an exchange algorithm to form a partition of length K and corresponding sets of parameters $\mathbf{b}_k$ such that the sum of the error sums of squares computed over all clusters is minimized:

$$Min\ Z = \sum_{k=1}^{K} \| \mathbf{X}^k\mathbf{b}_k - \mathbf{y}^k \|^2 . \tag{8}$$

Here, to guarantee the existence of a solution $\mathbf{b}_k$, it is required that the rank $\mathbf{X}^k = J$. A necessary condition for this is $I_k \geq J$ which implies $I \geq KJ$, where I_k is the number of observations/subjects in cluster k. Späth's methodology simultaneously solves for the optimal feasible partition $Q(K,I_k)$ and regression weights per cluster $\mathbf{b}_k$ that (locally) minimize expression (8). For the L_2 norm in expression (8), Späth (1982, 1985) has developed up and downdating formulae for the solution of these regression problems when an individual observation is added or removed utilizing QR-decompositions. His stepwise-optimal method works sequentially on the observations and is conceptually similar to K-means (MacQueen 1967). The original procedure can be summarized as follows:

1. Choose some initial partition $Q_1, \ldots, Q_K$ such that $|Q_k| \geq J$, and some starting observation;
2. Set $t = t + 1$ and reset $t = 1$ if $t > I$. For $i \in Q_j$ and $|Q_j| > I_k^*(I_k^* > J)$, examine whether there are clusters Q_k with $k \neq j$ such that shifting observation i from Q_j to Q_k reduces the objective function (expression (8)). If so, then choose Q_k such that the reduction becomes maximal and redefine $Q_j = Q_j - \{i\}$, $Q_k = Q_k + \{i\}$. Otherwise return to step 2.
3. Repeat step 2 as long as you get any reduction in the objective function; otherwise, stop.

One selects a solution with K^* clusters by choosing the solution with minimum value of Z in expression (8). According to Späth (1982), the final solution depends on the initial partition, on the starting observation, and on the choice of I_k^*, a minimum number of observations in each cluster. Because of problems with locally optimal solutions, Späth (1982) recommends running multiple analyses for a prespecified K, altering the initial starting partition and I_k^*.

The primary goal of this research is to extend the concept of cluster-wise regression to a stochastic context allowing for the possibility of fuzzy clusters, as well as mutually exclusive partitions. Given the documented problems with locally optimum solutions in Späth's (1985) deterministic procedure, we will devise a methodology that is hopefully less prone to such problems. In addition, we attempt to provide an AIC basis for selecting the most appropriate K^*.

3. Methodology

3.1. The Model

In addition to the notation developed prior to equation (1), let:

k = 1.....K clusters;
b_{jk} = the value of the j-th regression coefficient for the k-th cluster;
σ_k^2 = the variance term for the k-th cluster.

We assume y_i is distributed as a finite sum or mixture of conditional univariate normal densities:

$$y_i \sim \sum_{k=1}^{K} \lambda_k f_{ik} (y_i \mid X_{ij}, \sigma_k^2, b_{jk}) \tag{9}$$

$$= \sum_{k=1}^{K} \lambda_k (2\pi\sigma_k^2)^{-1/2} \exp \left[\frac{-(y_i - \mathbf{x}_i \mathbf{b}_k)^2}{2\sigma_k^2} \right], \tag{10}$$

where $\mathbf{X}_i = ((X_j))_i$ and $\mathbf{b}_k = ((b_j))_k$. That is, we assume an independent sample of subjects'/observations' dependent variable $y_1, y_2, \ldots, y_I$ drawn randomly from a mixture of conditional normal densities of underlying groups or clusters in unknown proportions $\lambda_1, \lambda_2, \ldots, \lambda_K$. Mixtures of univariate *unconditional* normal distributions have been the focus of many statisticians dating back to the seminal work by Pearson (1894) who derived estimators of the parameters of a mixture of two univariate normal distributors by equating sample moments to corresponding populations or theoretical moments involving the solution of a ninth degree polynomial equation. Charlier and Wicksell (1924) and Cohen (1967) simplified these computations considerably using method of moments estimators. Hasselblad (1966) was one of the first statisticians to formulate a maximum likelihood estimation scheme for mixture of two or more univariate normals.

Note that our mixture model is conceptually similar to the *unconditional* mixture approaches to pattern clustering originally proposed by Cooper

(1964), Wolfe (1965, 1967, 1970), and Day (1969), where $\mathbf{X}_i\mathbf{b}_k$ in expression (12) replaces the population mean/centroid μ_k (see also Ganesalingam and McLachlan 1981; McLachlan 1982; Sclove 1977; Symons 1981; Scott and Symons 1971; Marriott 1975; Hartigan 1975, pp. 113-124; and Basford and McLachlan 1985). In fact, expressions (9) and (10) generalize the Quandt (1972), Hosmer (1974), and Quandt and Ramsey (1978) stochastic switching regression models to more then two "regimes" (see also Veaux 1986). In addition, the estimation algorithm employed here differs from typical method of moments and moment generating function estimation approaches.

Given a sample of I independent subjects/observations, one can thus form a likelihood expression:

$$L = \prod_{i=1}^{I} \left[\sum_{k=1}^{K} \lambda_k (2\pi\sigma_k^2)^{-1/2} \exp \left[\frac{-(y_i - \mathbf{X}_i\mathbf{b}_k)^2}{2\sigma_k^2} \right] \right] \tag{11}$$

or

$$ln\, L = \sum_{i=1}^{I} ln \left[\sum_{k=1}^{K} \lambda_k (2\pi\sigma_k^2)^{-1/2} \exp \left[\frac{-(y_i - \mathbf{X}_i\mathbf{b}_k)^2}{2\sigma_k^2} \right] \right] . \tag{12}$$

Given K, $\mathbf{y}$, and $\mathbf{X}$, one wishes to estimate λ_k, σ_k^2, and b_{jk} in order to maximize L or $ln\, L$, where

$$0 \leq \lambda_k \leq 1 , \tag{13}$$

$$\sum_{k=1}^{K} \lambda_k = 1 , \tag{14}$$

$$\sigma_k^2 > 0 . \tag{15}$$

It is interesting to note several properties of this formulation. First, unlike finite mixtures of other types of density functions, the parameters of finite mixtures of normal densities are identified (see Yakowitz 1970; Yakowitz and Spragins 1968; and Teicher 1961, 1963). Second, there exist no sufficient estimators for the parameters of a normal mixture (Dynkin 1961). Third, unless (15) is imposed, consistent estimators are not possible given that the likelihood function is unbounded when $\sigma_k^2 = 0$. Finally, note that once estimates of λ_k, σ_k^2 and b_{jk} are obtained, one can assign each observation i to each cluster k (using Bayes rule) via the estimated posterior probability:

$$\hat{p}_{ik} = \frac{\hat{\lambda}_k f_{ik}(y_i \mid X_{ij}, \hat{\sigma}_k^2, \hat{b}_{jk})}{\sum_{k=1}^{K} \hat{\lambda}_k f_{ik}(y_i \mid X_{ij}, \hat{\sigma}_k^2, \hat{b}_{jk})} \tag{16}$$

This result renders a "fuzzy" clustering of the I subjects/observations. One could form partitions by applying the rule:

$$\text{Assign } i \text{ to } k \text{ iff } \hat{p}_{ik} > \hat{p}_{il} \quad \text{for all} \quad l \neq k = l...K .$$

3.2. The Algorithm

The maximum likelihood estimates of λ_k, $\mathbf{b}_k$, σ_k^2 and p_{ik} are found by initially forming an augmented log likelihood function to reflect the λ_k constraints in expression (14):

$$\Phi = \sum_{i=1}^{I} ln \left[\sum_{k=1}^{K} \lambda_k f_{ik}(y_i \mid X_{ij}, \sigma_k^2, b_{jk}) \right] - \mu(\sum_k \lambda_k - 1) . \tag{17}$$

The resulting maximum likelihood stationary equations are obtained by equating the first order partial derivatives of the augmented log likelihood function in (17) to zero:

$$\frac{\partial \Phi}{\partial \lambda_k} = \sum_{i=1}^{I} \frac{1}{\sum_k \lambda_k f_{ik}(*)} f_{ik}(*) - \mu = 0 \tag{18}$$

$$\frac{\partial \Phi}{\partial \sigma_k^2} = \sum_{i=1}^{I} \frac{1}{\sum_k \lambda_k f_{ik}(*)} \lambda_k \frac{\partial f_{ik}(*)}{\partial \sigma_k^2} = 0 \tag{19}$$

$$\frac{\partial \Phi}{\partial b_{jk}} = \sum_{i=1}^{I} \frac{1}{\sum_k \lambda_k f_{ik}(*)} \lambda_k \frac{\partial f_{ik}(*)}{\partial b_{jk}} = 0 , \tag{20}$$

where $f_{ik}(*)$ is used for $f_{ik}(y_i \mid X_{ij}, \sigma_k^2, b_{jk})$. To estimate μ, we multiply both sides of equation (18) by λ_k and then sum both sides over k:

$$\sum_{i=1}^{I} \frac{\sum_k \lambda_k f_{ik}(*)}{\sum_k \lambda_k f_{ik}(*)} - \mu \sum_k \lambda_k = 0$$

or

$$\hat{\mu} = I . \tag{21}$$

To estimate λ_k, we multiply both sides of equation (18) by λ_k and simplify:

$$\sum_{i=1}^{I} \frac{\lambda_k f_{ik}(*)}{\sum_k \lambda_k f_{ik}(*)} - \lambda_k \mu = 0 , \tag{22}$$

or

$$\sum_{i=1}^{I} \hat{p}_{ik} - \lambda_k I = 0 , \tag{23}$$

and

$$\hat{\lambda}_k = \frac{\sum_{i=1}^{I} \hat{p}_{ik}}{I} . \tag{24}$$

In order to estimate σ_k^2 and b_{jk}, we use the definition of $\hat{p}_{ik}$ in (16) and re-express (19) and (20) as:

$$\frac{\partial \Phi}{\partial \sigma_k^2} = \sum_{i=1}^{I} \hat{p}_{ik} \frac{\partial \log f_{ik}(*)}{\partial \sigma_k^2} = 0 \tag{25}$$

$$\frac{\partial \Phi}{\partial b_{jk}} = \sum_{i=1}^{I} \hat{p}_{ik} \frac{\partial \log f_{ik}(*)}{\partial b_{jk}} = 0 . \tag{26}$$

Thus, the maximum likelihood equations for estimating the parameters σ_k^2 and b_{jk} are weighted averages of the maximum likelihood equations $\frac{\partial \log f_{ik}(*)}{\partial \theta} = 0$, where θ reflects the parameter of interest, arising from each component separately and the weights are the posterior probabilities of membership of the subjects/observations in each cluster. This particular structure gainfully lends itself to the development of a two stage E-M algorithm (Dempster, Laird, and Rubin 1977) for the estimation of these parameters (see Hosmer 1974; Veaux 1986). In the E-stage, one estimates λ_k and p_{ik} via expression (16) and (24). In the M-stage, one estimates b_{jk} and σ_k^2 via K weighted least squares regressions. In order to show this M-stage, we expand (25) and (26):

$$\frac{\partial \Phi}{\partial \mathbf{b}_k} = \sum_{i=1}^{I} \frac{1}{\Sigma \lambda_k f_{ik}(*)} \cdot \lambda_k (2\pi\sigma_k^2)^{-1/2} \times$$

$$\exp\left[\frac{-(y_i - \mathbf{X}_i\mathbf{b}_k)^2}{2\sigma_k^2}\right] \cdot \frac{2(y_i - \mathbf{X}_i\mathbf{b}_k)\mathbf{X}_i}{2\sigma_k^2} = \mathbf{0}$$

$$= \sum_{i=1}^{I} \hat{p}_{ik}\,(y_i - \mathbf{X}_i\mathbf{b}_k)\mathbf{X}_i = \mathbf{0}\,, \tag{27}$$

which are identical to the stationary equations derived by solving the weighted least squares problem where $\mathbf{y}$ and $\mathbf{X}$ are each weighted by $\hat{p}_{ik}^{1/2}$. Thus, the entire set of $\mathbf{b}_k$ is derived by performing K separate weighted least-squares analyses. Once this is done, the estimates of σ_k^2 follow:

$$\frac{\partial \Phi}{\partial \sigma_k^2} = \sum_{i=1}^{I} \frac{1}{\sum_k \lambda_k f_{ik}(*)} \left[\lambda_k \exp\left[\frac{-(y_i - \mathbf{X}_i\mathbf{b}_k)^2}{2\sigma_k^2}\right](-1/2(2\pi\sigma_k^2)^{-3/2}2\pi)\right.$$

$$\left. + \lambda_k (2\pi\sigma_k^2)^{-1/2} \exp\left[\frac{-(y_i - \mathbf{X}_i\mathbf{b}_k)^2}{2\sigma_k^2}\right] \frac{1/2(y_i - \mathbf{X}_i\mathbf{b}_k)^2}{\sigma_k^4}\right] = 0$$

$$= \sum_{i=1}^{I} \hat{p}_{ik}\left[\frac{-1}{2\sigma_k^2} + \frac{(y_i - \mathbf{X}_i\mathbf{b}_k)^2}{2\sigma_k^4}\right] = 0\,. \tag{28}$$

Multiplying both sides of (28) by $2\sigma_k$ and simplifying, one obtains:

$$\hat{\sigma}_k^2 = \frac{\sum_{i=1}^{I} \hat{p}_{ik}(y_i - \mathbf{X}_i\mathbf{b}_k)^2}{\sum_{i=1}^{I} \hat{p}_{ik}}\,. \tag{29}$$

Thus, $\hat{\sigma}_k^2$ can be obtained during the K weighted least-squares procedures for estimating $\mathbf{b}_k$. Note, because (17) becomes unbounded as $\sigma_k^2 \to 0$, $\hat{\sigma}_k^2$ is set to a default small positive value (.01) if it becomes small during these iterations.

Thus, the computation of the maximum likelihood estimates is facilitated by the use of this E-M algorithm. For given starting values of the parameters, the expectation (E phase) and maximization (M phase) steps of this algorithm are alternated until convergence of a sequence of log likelihood values is obtained. Dempster, Laird, and Rubin (1977) prove that:

$$\Phi(\Theta^{(m+1)}) \geq \Phi(\Theta^{(m)}), \tag{30}$$

where m is the iteration counter, indicating that the E-M algorithm provides monotone increasing values of the objective function. Given the constraint $\sigma_k^2 \geq .01$, one can show that Φ is bounded from above and convergence to at least a local maximum can be established (cf. Titterington, Smith, and Makov 1985). While several authors (e.g., Everitt and Hand 1981 and Redner and Walker 1984) have documented the potentially slow convergence rate of E-M procedures for estimating the parameters of unconditional mixture distributions, we find that our E-M procedure typically converges in 100 or less iterations. Moreover, the iterations are processed much faster than if a gradient based procedure had been used. Acceleration procedures discussed by Peters and Walker (1978), Wilson and Sargent (1979), and Louis (1982) are currently being investigated. We provide a Monte Carlo analysis in the next major section to investigate the performance of this E-M algorithm in a reasonably rigorous manner.

Our approach to identify the appropriate number of clusters K^* in such mixture clustering procedures (cf. Sclove 1977) involves the use of the Akaike Information Criteria (Akaike 1974) which is defined as:

$$AIC(K) = -2\ ln[max\ L(K)] + 2n(K), \tag{31}$$

where $n(K)$ is the effective number of parameters estimated in a K clusterwise regression solution:

$$n(K) = JK + 2K - 1. \tag{32}$$

This AIC criteria has been previously used in an unconditional mixture/clustering context by Sclove (1983). However, as pointed out by Bozdogan (1983) and Sclove (1987), one major problem with the use of such a criterion is that the theoretical justification for use of AIC relies on the same conditions as the usual asymptotic theory of the GLR test. In this context, some analytical conditions required for series expansions yielding the AIC are not strictly met (see McLachlan and Basford 1988, p. 28; Sclove 1987), and the criteria can be thus regarded as "heuristic figures of merit" where one selects K^* which renders minimum $AIC(K)$.

Note that the likelihood ratio criterion for testing the hypothesis of K_1 versus K_2 clusters, where $K_1 < K_2$, does not have its usual asymptotic distribution as mentioned by Hartigan (1977), Binder (1978), and McLachlan and Basford (1988, p. 27). Basford and McLachlan (1985) have adapted Wolfe's (1971) approach in introducing a constant to improve a X^2 approximation for the likelihood ratio test. However, the reliability of such as approximation

will naturally depend on the size of I. McLachlan (1987) has recently examined the boot-strapping of the log likelihood ratio statistic to assess the null distribution of $-2\log L$. Further research is required in this area.

One of the appealing properties of maximum likelihood estimators is that, under typical regularity conditions, these estimators are asymptotically normal. Define $\mathbf{b}$ as a vector of all the $(b_1, b_2, \ldots, b_K)$ estimated coefficients in a maximum likelihood context, and $\mathbf{B}$ as the corresponding vector of unknown population parameters $(B_1, B_2, \ldots, B_K)$. Then, according to Theil (1971),

$$\sqrt{I}\,(\mathbf{b} - \mathbf{B}) \xrightarrow{d} N(\mathbf{O}, \lim(R(\mathbf{B}) / I)^{-1}) \tag{33}$$

where:

$$R(\mathbf{B}) = -E\left[\frac{\partial^2 \Phi}{\partial \mathbf{B} \partial \mathbf{B}'}\right], \tag{34}$$

the information matrix. According to Judge, Griffiths, Hill, Lütkepohl, and Lee (1985), replacing $\lim(R(\mathbf{B}) / I)$ by a consistent estimator does not change the asymptotic distribution of the test statistics or confidence intervals for $\mathbf{b}$. Here, the consistent estimator utilized is:

$$\mathbf{F} = \frac{1}{I}\left[\sum_{i=1}^{I}\left[\frac{\partial \Phi}{\partial \mathbf{b}^*}\right]\left[\frac{\partial \Phi}{\partial \mathbf{b}^*}\right]'\right]_{\mathbf{b}^* = \mathbf{b}}, \tag{35}$$

and the asymptotic variances of $\mathbf{b}$ can be defined as the main diagonal elements of $\mathbf{F}^{-1}$, the asymptotic variance covariance matrix. From (33) - (35), it follows that an asymptotic $(1 - \alpha)$ 100% confidence interval for B_n is given by

$$(b_n - Z_{\alpha/2}\sqrt{f_{nn}^{-1}},\ b_n + z_{\alpha/2}\sqrt{f_{nn}^{-1}}), \tag{36}$$

where $Z_{\alpha/2}$ is the central value of a normal distribution with mean zero and variance one and f_{nn}^{-1} is the asymptotic estimate of the variance of b_n.

3.3. Synthetic Data Analysis

The synthetic data in Table 1 were analyzed by our conditional mixture E-M based procedure. Table 2 presents a statistical and computational summary for $K = 1$ to 4 clusters. As clearly delineated in this table, the $K = 2$ cluster solution is the "best" one given that the minimum AIC is obtained here. The recovered parameters are also shown in the table for this small

TABLE 2

Conditional Mixture Maximum Likelihood
Procedure Results for Synthetic Data

K	Number of Iterations Required for Convergence	ln L	AIC
1	2	-39.70	85.39
2	6	-12.86	39.73*
3	7	-12.86	47.73
4	8	-12.86	55.73

*minimum AIC

Recovered Parameters:

$\underset{\sim}{\lambda} = (.5 \quad .5)$

$\underset{\sim}{o} = (.5 \quad .5)$

$\underset{\sim}{b} = \begin{pmatrix} 1 & -1 \\ 2 & -2 \end{pmatrix}$

$\underset{\sim}{P} = \begin{matrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{matrix}$

illustrative data set. We performed 20 additional computer runs for the two cluster solution, varying the starting estimates of $\mathbf{b}_k$ using the uniform distribution $U(-2,2)$. The starting values produced log likelihoods in the range of -39.13 to -61.32. In all 20 cases, the procedure converged within 5-7 iterations to this same globally optimum solution presented in Table 2. Given that the starting values were generated from the same distribution as the actual values, we performed an additional 20 computer runs utilizing $U(-20,20)$ for the starting $\mathbf{b}_k$. This generated initial log likelihood values ranging from -69.95 to -432.83. Here, we recovered the actual parameter values shown in Table 2 in 18 of the 20 cases, showing some deterioration in performance as the "quality" of the starting values deteriorated. Note that Späth's (1985) procedure was able to recover the actual $\mathbf{b}_k$ values in 14 of 20 computer runs in the version of the computer program we initially purchased from him.

TABLE 3

Independent Factors and Levels Utilized in the Monte Carlo Analysis

FACTOR	LEVELS	CODE
A. Number of Clusters (K)	K=2	2
	K=3	3
	K=4	4
B. Number of Independent Variables (J) in $\underset{\sim}{X}$	J=2	2
	J=5	5
	J=8	8
C. Number of Observations (I)	I=50	50
	I=100	100
	I=150	150
D. Difference in Scale of σ_k e.g., for K=3	$\underset{\sim}{\sigma}$ = 1,2,3	1
	$\underset{\sim}{\sigma}$ = 2,4,6	2
	$\underset{\sim}{\sigma}$ = 3,6,9	3
E. Comparative Range of Mixing Proportions (λ_k)	Equal λ's	1
	Unequal λ's	2
F. Distribution of each b_{jk}	$b_{jk} \sim N(0,1)$	1
	$b_{jk} \sim N(k,1)$	2
	$b_{jk} \sim N(2k,1)$	3
G. Estimation Option for σ_k^2	σ_k^2 fixed at true values (1)	1
	estimate σ_k^2	2

4. Monte Carlo Analysis

In order to examine systematically the performance of the conditional mixture E-M algorithm, a Monte Carlo analysis was performed where some seven factors were experimentally varied: K,J,I, scale of σ_k^2, λ_k, the distribution of $\mathbf{b}_k$, and σ_k^2 estimation options. Table 3 describes these seven independent factors and the various levels tested for each factor. These seven factors

TABLE 4

$3^5 2^2$ Asymmetric Fractional Factorial Design

TRIAL	A	B	C	D	E	F	G
1	2	2	50	1	1	1	1
2	2	2	50	1	2	2	1
3	2	2	50	1	1	3	2
4	2	5	100	3	1	1	1
5	2	5	100	3	1	2	2
6	2	5	100	3	2	3	1
7	2	8	150	2	2	1	2
8	2	8	150	2	1	2	1
9	2	8	150	2	1	3	1
10	3	2	100	2	2	2	2
11	3	2	100	2	1	3	1
12	3	2	100	2	1	1	1
13	3	5	150	1	1	2	1
14	3	5	150	1	2	3	1
15	3	5	150	1	1	1	2
16	3	8	50	3	1	2	1
17	3	8	50	3	1	3	2
18	3	8	50	3	2	1	1
19	4	2	150	3	1	3	1
20	4	2	150	3	1	1	2
21	4	2	150	3	2	2	1
22	4	5	50	2	2	3	2
23	4	5	50	2	1	1	1
24	4	5	50	2	1	2	1
25	4	8	100	1	1	3	1
26	4	8	100	1	2	1	1
27	4	8	100	1	1	2	2

were combined via an asymmetric fractional factorial design (cf. Addelman 1962) for main effects only estimation. Twenty-seven experimental trials were devised where the seven factors were varied according to the $3^5 2^2$ fractional factorial design portrayed in Table 4. Such procedures have been previously utilized by DeSarbo (1982) and DeSarbo and Carroll (1985) in the psychometric literature for preliminary testing of new algorithms. Note that each trial (or row) of the experimental design defined a specific level for each of the seven factors listed in Table 3. Based on the stipulated levels of J and I, $\mathbf{X}$ was randomly generated from a uniform distribution. Given designated levels of σ_k^2, λ_k, and $\mathbf{b}_k$, $\mathbf{y}$ was generated via the mixture specification described in equation (9), and the conditional maximum likelihood cluster-wise linear regression procedure was executed.

The dependent/performance measures collected were:

1. **Number of iterations required for convergence.** This was to measure the computational effort involved in processing given CPU time was unavailable. Note, a maximum of 100 iterations was specified.

2. $AIC_A - AIC_R$. This is the difference in the Akaike information statistic between that obtained for the actual values generated synthetically for the Monte Carlo analysis (AIC_A) and that obtained via the solution recovered by the methodology (AIC_R). Note that this difference was taken to eliminate the dependence of such a measure on I, J and K. Positive values for this difference would indicate the methodology recovered a better solution than compared to the actual parameters.
3. b_k **parameter recovery.** A root mean square between the actual b_k and estimated $\hat{b}_k$ (after appropriate permutation) is calculated to measure how well the procedure can recover the clusters' regression coefficients.
4. σ_k **parameter recovery.** A root mean square between the actual σ_k and estimated $\hat{\sigma}_k$ (after appropriate permutation) is calculated to measure how well the procedure can recover these parameters.
5. p_{ik} **recovery.** A root mean square between these actual posterior probabilities and estimated $\hat{p}_{ik}$ is calculated (after appropriate permutation) to measure how well the procedure can reproduce these cluster membership probabilities.
6. λ_k **recovery.** Finally, a root mean square between the actual mixing proportions and estimated $\hat{\lambda}_k$ is calculated (after appropriate permutation) to measure how well the procedure can recover these mixing proportions.

Thus, these six measures encompass the three major areas of algorithm performance: computational demands, data reproduction, and parameter recovery.

Table 5 presents the results for these six dependent measures for each of the twenty-seven trials designated by the asymmetric fractional factorial design presented in Table 4. As can be seen in Table 5, only two of the twenty-seven trials (#18, #19) failed to reach convergence within the maximum limit of 100 iterations. Also, note that all twenty-seven trials resulted in positive values for the second dependent measure indicating the procedure *always* provided estimates whose resultant AIC statistic was better (lower) than that produced by the actual parameters. the difference in the magnitudes between the four RMS dependent measures 3, 4, 5, and 6 reflect the differences in the scale in the numbers utilized as parameter values rather than better/poorer fits. Table 5 also presents the correlations between these six measures. Of particular note is the rather large positive correlation (0.764) between dependent measures #3 and #4 indicating an association in attempting to recover the parameters (b_k, σ_k^2) of the conditional normal distributions. Note, in none of the trials were there estimates of σ_k^2 near 0.00.

TABLE 5

Dependent Measure Results from the Monte Carlo Analysis

DEPENDENT MEASURE:

TRIAL:	1	2	3	4	5	6
1	38	3.602	1.242	1.000	.559	.051
2	10	5.029	2.361	1.000	.744	.229
3	14	7.682	.468	.404	.221	.137
4	33	6.720	.810	.000	.085	.049
5	52	7.511	.944	.298	.153	.017
6	20	6.971	3.069	3.000	.610	.382
7	41	17.137	.698	.432	.248	.134
8	32	15.715	2.489	2.000	.647	.051
9	10	26.284	3.476	2.000	.761	.051
10	37	5.465	1.918	4.031	.328	.235
11	51	10.761	3.161	2.828	.379	.132
12	37	4.686	1.652	3.266	.346	.078
13	34	8.910	.900	.816	.266	.023
14	40	9.826	4.231	1.414	.606	.171
15	84	2.169	1.662	.604	.425	.259
16	55	26.423	4.560	4.899	.452	.273
17	20	34.840	6.917	4.714	.480	.156
18	100	30.223	4.892	4.243	.486	.265
19	100	10.832	8.497	5.612	.372	.206
20	57	6.254	2.424	2.655	.247	.171
21	64	11.633	6.551	3.674	.374	.232
22	15	40.187	5.806	4.345	.337	.210
23	29	36.779	7.729	4.000	.399	.170
24	56	23.237	3.801	2.828	.313	.147
25	62	14.230	3.978	.707	.412	.094
26	94	54.187	4.200	1.225	.500	.250
27	44	95.596	3.173	1.576	.473	.059

CORRELATIONS:

	1	2	3	4	5	6
1	1.000	.075	.273	.170	-.165	.273
2		1.000	.354	.163	.161	-.025
3			1.000	.764	.241	.395
4				1.000	.121	.480
5					1.000	.245
6						1.000

Table 6 presents the results of the six regression analyses performed, one for each of the dependent measures. Here, as in conjoint analysis (cf. Green and Rao 1971), the design matrix is converted into dummy variables prior to the regression analysis. Such a methodology has been similarly used in De Soete, DeSarbo, Furnas, and Carroll (1984), DeSarbo and Carroll (1985), and DeSarbo (1982) in the Monte Carlo testing of new methodologies.

Estimating solutions with larger numbers of clusters significantly increases the number of iterations required for convergence. While not sig-

TABLE 6

Regression Analyses of Monte Carlo Results

	DEPENDENT MEASURE:					
	1	2	3	4	5	6
INTERCEPT	34.06	1.33	0.96	0.38	0.48	0.10
K=3	23.11*	4.07	1.59*	1.85**	-0.03	0.05
K=4	30.11*	21.81**	3.40**	1.83**	-0.07	0.05
J=5	-5.00	8.49	0.08	-0.80	-0.04	-0.01
J=8	5.56	27.63**	0.68	-0.30	0.10	-0.02
I=100	10.33	-0.21	-1.65*	-1.17*	-0.08	-0.04
I=150	13.89	-11.03	-0.76	-0.91*	-0.01	-0.04
σ of 2	-12.44	-2.33	0.95	1.89**	-0.05	-0.01
σ of 3	9.00	-6.65	1.83*	2.26**	-0.11	0.05
UNEQUAL λ_k	1.89	1.06	0.53	0.36	0.08	0.12**
b_{jk}~N(k,1)	-14.33	4.20	0.15	0.41	0.05	-0.02
b_{jk}~N(2k,1)	-20.11	-0.02	1.59*	0.84	0.10	0.01
σ_k^2 ESTIMATED	-7.61	7.09	-1.09	-0.36	-0.14	-0.01
S.E.	22.74	15.21	1.32	0.85	0.15	0.08
R^2	0.59	0.69	0.81	0.86	0.55	0.59
adj R^2	0.23	0.43	0.65	0.74	0.16	0.24
F	1.66	2.65*	4.96**	7.10**	1.41	1.68

* $p \leq .05$

** $p \leq .01$

nificant, data with larger J and larger I also tend to increase the number of iterations. Thus, larger data sets and solutions estimating larger numbers of parameters tend to increase computational demands, although the regression equation as a whole is not significant.

A somewhat surprising finding is seen with respect to the second dependent measure concerning the difference in AIC statistics produced by the actual parameters vs. recovered ones. Here, as the number of estimated parameters increase ($K = 4, J = 8$), the procedure is somewhat more likely to recover parameters with associated lower AIC statistics than compared to that produced by the actual parameters. Thus, as the dimensionality of the parameter space increases, all else held equal, there is a greater chance of finding a better solution. Note that this regression equation is significant at $p \leq .05$.

Some unanticipated results are found concerning the regression analysis conducted with dependent measures #3 involving $\mathbf{b}_k$ recovery. While the result that solutions involving larger number of clusters (and more $\mathbf{b}_k$ parameters) tend to detract from recovery and larger sample sizes enhance $\mathbf{b}_k$ recovery make intuitive sense, the positive and significant impact of the larger σ_k^2 scale and $b_{jk} \sim N(2K,1)$ levels are a bit harder to interpret.

Presumedly, as σ_k^2 gets larger, the variance of b_k increases rendering larger errors in recovery. Note that this regression equation is quite significant.

Some similar patterns are also seen with respect to regression analysis performed on dependent measure #4, σ_k recovery. Here, σ_k recovery is reduced as the number of σ_k terms increases and as the scale and difference of the σ_k increase. All else held equal, however, as the sample size increases, it becomes significantly easier to recover the σ_k values. This makes considerable sense in light of traditional statistical estimation theory concerning the impact of higher degrees of freedom in estimation. Again, this regression equation is also significant at $p \leq .01$.

The final two regression equations for dependent measures #5 (p_{ik} recovery) and #6 (λ_k recovery) are not significant. For the p_{ik} RMS equation, no factor level is significant. For the λ_k RMS equation, it appears that estimating unequal λ_k tends to detract from λ_k recovery; however, the regression equation again is not significant.

Thus, the Monte Carlo analysis appears to result in several rather interesting findings. As the number of parameters to be estimated increases, all else held equal, computational time will increase and parameter recovery, in general, will suffer as is the case in most nonlinear estimation problems. Similarly, increasing the sample size, holding all else equal, may also increase computational demands, but will typically improve parameter recovery. Finally, increasing the variance of the parameters to be estimated will also tend to result in poorer parameter recovery. However, given the preliminary nature of these analyses, these results must be subject to further testing.

Some obvious limitations of this Monte Carlo analysis must be noted. The use of the fractional factorial design does not allow the flexibility of measuring possible interaction effects between these factors studied in the analysis. Clearly, assuming computational time/expense was not a limitation, a full factorial design would have been a more comprehensive design to use in order to estimate possible significant higher order interaction terms. In addition, the design should have been replicated in order to improve the degrees of freedom for estimation. Finally, more levels for each of the factors should be investigated, and other factors (e.g., cluster size and shape) introduced in the design. We leave these projects for future research.

5. Application — Trade Show Performance

5.1. Study Description

Trade shows are promotional events used by marketers to draw a large number of prospective buyers to view exhibits of products/services in a few concentrated days. Such trade shows have become a very popular medium

for promoting products and services, especially in the industrial sector. Cleaver (1982) published figures indicating that over 91,000 firms display such exhibits at some 8,000 trade shows to over 31 million prospective buyers at a total cost of $7 billion annually. Many firms will allocate up to 25% of their total promotion budget for trade shows (Mee 1983a). Historically, trade show participation has been viewed as an extension of a firm's personal selling effort. However, Bonoma (1983) revealed that trade shows have a much broader role than merely generating sales. Many firms consider such non-selling factors as image enhancement, gathering competitive information, and improving corporate morale as equal to, if not more important, than identifying leads on making sales.

Recently, Kerin and Cron (1987) conducted a survey of trade show exhibit managers and senior marketing executives in 129 firms that were heavy participants in trade shows. One of their objectives was to investigate the selling vs. non-selling role of trade shows. A self-administered questionnaire was separately mailed to the trade show exhibit manager and the senior marketing executive in each firm. We will purposely focus on the latter questionnaire sent to the senior marketing executive since it focused on perceptions of trade show performance and various marketing-related variables identified in the literature as affecting such perceptions. These marketing executives were asked to rate the firm's trade show performance on some eight functions documented in the literature (see Haas 1982; Bonoma 1983; Hutt and Speh 1985; Dunn and Barban 1986):

1. Identifying New Prospects
2. Servicing Current Customers
3. Introducing New Products
4. Selling at the Trade Show
5. Enhancing Corporate Image
6. Testing of New Products
7. Enhancing Corporate Morale
8. Gathering Competitive Information

Overall trade show performance was rated also. Each of these performance aspects was rated on a 7-point Likert type scale (1=very poor; 7=very good) which we shall treat as metric scales (cf. Guilford 1954, pp. 15-16; and Green and Tull 1978). In addition, data on a number of individual difference items were collected (we will describe these later). Kerin and Cron (1987) had performed a factor analysis on the eight performance functions listed above and uncovered two dimensions accounting for 59.1% of the variance, roughly corresponding to the selling and non-selling roles of trade shows conceptually identified by Bonoma (1983). Our investigation will examine a multiattribute analysis of the performance function data where we shall use overall trade

TABLE 7

Total Sample Regression Results on Trade Show Performance Data

INTERCEPT	3.03
X_1	0.15***
X_2	-0.02
X_3	0.09
X_4	-0.04
X_5	0.09
X_6	0.18***
X_7	0.07
X_8	0.04
S.E.	0.85
R^2	0.37
adj R^2	0.33
S.S.E.	87.67
F	8.87***
I	129

* $p \leq .10$
** $p \leq .05$
*** $p \leq .01$

show performance as the dependent variable and the data on the eight performance functions listed above as the independent variables. Our goal is to examine whether groups of firms evaluate overall trade show performance differently in terms of these eight aspects, and if so, estimate their different regression coefficients and group membership probabilities via the new clusterwise-linear regression methodology discussed.

5.2. Preliminary Analyses

We will analyze the data for overall performance (y) and the eight performance functions ($X_1, \ldots, X_8$) for these 129 firms. Treating all 129 executives as members of one large cluster or group, Table 7 presents the resulting regression analysis of regressing overall performance on the eight performance functions. As can be seen, identifying new prospects (X_1) and new product testing (X_6) appears to be most significantly related to evaluations of overall trade show performance. Thus, for the entire sample, it appears that these two selling-related aspects dominate the analysis for the entire sample. The issue remaining is whether there exist distinct groups of firms which exhibit different regression coefficients.

In order to address this research issue of group regression coefficients, we initially applied Späth's (1982, 1985) clusterwise linear regression procedure. We ran 20 trials for each solution from $K = 2$ to 5 and utilized

Späth's minimum objective function rule to select both the number of clusters and the particular solution. The $K = 2$ cluster solution was selected using Späth's minimum objective function rule. Table 8 contains the best $K = 2$ cluster solution obtained from these analyses. This table presents multiple regression analyses and corresponding asymptotic significance tests or each of the two derived groups. The b_k coefficients are identical to those obtained from Späth's procedure, but significance tests are missing from Späth's procedure since it is deterministic. While it is not good practice to consider these significance tests appropriate (since the data were initially utilized to form the groups), we merely present them as "heuristic figures of merit" in order to gain some insight into the structure of the data as derived from this alternative methodology. The first cluster of some 72 executives appear to derive their overall performance evaluation on primarily non-selling functions such as enhancing corporate image and morale (X_5 and X_7) and new product introduction and testing (X_3 and X_6). Note the significant negative coefficient on selling at trade shows (X_4). This is a cluster of marketing executives who appear to stress particular non-selling and new products aspects of their trade shows. The second cluster, however, is not as clearly interpretable. Here, identifying prospects (X_1) and new product testing (X_6) are the most significant functions impacting on overall trade show performance evaluation, although these relationships are not as strong as those reported in the previous cluster. As such, this second cluster of 57 marketing executives appears to resemble the total group structure as reported in Table 6.

5.3. Conditional Mixture Maximum Likelihood Procedure Results

Our conditional normal mixture maximum likelihood methodology was applied to these data for $K = 1$ to 4 clusters. Table 9 presents the number of iterations required for convergence, *ln L*, and AIC statistics for each solution. According to the minimum AIC rule, the $K = 2$ cluster solution appears to be the best one and will thus be reported here. Table 10 presents a summary of the various paramcter values and statistics for this two cluster solution. Cluster one, composed of 59 marketing executives, evaluates trade shows primarily in terms of evaluations on non-selling dimensions including servicing new customers (X_2) and enhancing corporate image and morale (X_5 and X_7). Note the significant negative coefficients for introducing new products (X_3) and selling at trade shows (X_4) which also substantiates this non-selling orientation. The second cluster of 70 marketing executives appears quite different than this first cluster. Here, identifying new prospects (X_1), introducing new products (X_3), selling at trade shows (X_4), and new product testing (X_6) are highly significant. The significant negative coefficient on servicing current customers (X_2) also helps substantiate this "selling" orientation.

TABLE 8

Späth's Clusterwise Linear Regression Two Cluster Solution - Multiple Regression Analyses

	CLUSTER 1	CLUSTER 2
INTERCEPT	2.27	3.72
X_1	0.09	0.18*
X_2	-0.00	-0.02
X_3	0.15**	-0.06
X_4	-0.17***	0.04
X_5	0.21***	0.04
X_6	0.27***	0.15 **
X_7	0.09*	0.00
X_8	0.04	0.11
S.E.	0.73	0.94
R^2	0.56	0.31
adj.R^2	0.51	0.20
S.S.E.	33.91	41.99
F	10.07***	2.75**
I	72	57

* $p \leq .10$
** $p \leq .05$
*** $p \leq .01$

TABLE 9

Conditional Mixture Maximum Likelihood Procedure Results for K=1-4 Clusters

K	Number of Iterations Required for Convergence	ln L	AIC
1	2	-158.1	336.3
2	32	-141.6	325.2*
3	62	-132.7	329.4
4	46	-130.9	347.8

* Minimum AIC Solution K=2

This solution appears to be more congruent with previous literature (cf. Bonoma 1983; Haas 1982; Hutt and Speh 1985; Dunn and Barban 1986) and the empirical results reported in Kerin and Cron (1987) than does the Späth two-cluster solution. In addition, the effects are stronger here than in Späth's solution producing higher adjusted R^2's when placed in a deterministic regression context. In fact, this conditional normal mixture maximum likelihood solution obtains a lower Späth objective function (expression 8) value than the Späth $K = 2$ solution! Table 11 presents a cross classification of membership for the Späth and conditional mixture E-M based procedure. As shown, only 68 of the 129 executives are classified similarly. The resulting phi coefficient calculated from this table is only 0.065 indicating little association between the two classifications. The Späth solution produced a log likelihood value of −156.99 when substituted in the conditional mixture E-M based procedure as compared to −141.58 for the solution reported earlier in Table 10. Using the Späth solution as an initial starting solution for the conditional mixture based E-M procedure produced (after 11 iterations) a solution with a log likelihood value of −142.53, whose $\mathbf{b}_k$ values had correlations with those in Table 10 of .996 and .983, and whose λ and σ^2 values differed by .01 and .02 respectively. At any rate, it is interesting to see how running a total group analysis such as reported in Table 7 can mask the true structure in a set of data.

Having identified two schemes used by marketing managers to evaluate their overall trade show performance, the usefulness of our classification was evaluated by attempting to describe the factors distinguishing between the two groups. Much of what has been written concerning trade show management is descriptive of the experiences of managers involved in aspects of trade show management (e.g., Cavanaugh 1976; Hatch 1981; Konikow 1983; Rich 1985). A number of studies that are descriptive of trade show management have been supported by the National Trade Show Bureau (Mee 1983a, 1983b, 1984). Perhaps the first effort to systematically analyze trade show management was Lilien's (1983) research on trade show budgeting and participation. This study identified factors related to how much an individual firm spent on trade shows and to which shows the firm participated. The research by Kerin and Cron (1987) on the determinants of high trade show performance evaluations also provides a good framework for identifying factors related to trade show performance.

Based on this review of the literature and interviews with marketing and exhibit managers, a list of factors were derived which are potentially related to whether marketing managers evaluate their trade show performance primarily on selling or non-selling dimensions. A complete list of the individual difference items collected in the Kerin and Cron (1987) study along with a description of their measurement are provided in Table 12. The variables

TABLE 10

Conditional Mixture Maximum Likelihood K=2 Cluster Solution

	CLUSTER 1	CLUSTER 2
INTERCEPT	4.093***	2.218***
X_1	0.126	0.242***
X_2	0.287***	-0.164***
X_3	-0.157**	0.206**
X_4	-0.133***	0.074**
X_5	0.128*	0.072
X_6	0.107	0.282***
X_7	0.155**	-0.026
X_8	-0.124	0.023
R^2	0.73	0.76
adj.R^2	0.69	0.73
S.S.E.	20.37	12.98
I	59	70
λ_k	0.489	0.511
σ_k	0.589	0.504

* $p \leq .10$
** $p \leq .05$
*** $p \leq .01$

TABLE 11

Membership Comparisons for Trade Show Data Analyses

		Späth's Cluster: 1	2	Totals
Conditional Mixture, E-M Procedure's Cluster:	1	35	24	59
	2	37	33	70
	Totals	72	57	129

TABLE 12

Independent Variables for Profiling Evaluation Groups

Variable	Description
INDUSTRY INFLUENCES:	
Stage of industry life cycle	Five point scale: introduction, growth, early maturity, maturity, and decline.
Degree of product customization	Percent of sales in customized products.
Major industry group	Percent of sales in each of the following: raw materials, component parts, major capital equipment, operating supplies, consumer durables, consumer nondurables, ad services
COMPANY INFLUENCES:	
Annual sales volume	Dollar figure
Number of direct customers	Number
Sales concentration	Percent of sales to top ten customers
Technical complexity	Five point Likert scale (1 = technically simple to 5 = technically complex).
Trade show budgeting	Percent of sales promotion budget spent on trade shows
Importance to top management	Five point Likert scale (1 = Not important, 5 = Very important)
Sales growth	Last year's percent
TRADE SHOW INFLUENCES:	
Written objectives[a]	Existence of formal written objectives for overall trade show effort
MARKETING MANAGER INFLUENCES:	
Length of time in position	Years in present position
Involvement in show decisions	A summative index including involvement in budgeting, policies, evaluation, setting objectives, participation, and working with exhibit manager on a five point Likert scale (1 = minimally involved to 5 = extensively involved).[b]

[a]Exhibit manager is the key informant for this variable, while the marketing manager provided information on the remaining variables.

[b]The Cronback alpha for this measure was .87.

are organized into a framework similar to that used by Kerin and Cron (1987), which consists of (a) industry influences, (b) company influences, and (c) trade show strategy influences. In addition, a fourth set of influencing factors were considered in this study and are referred to as the marketing manager's influences. This was considered to be appropriate because the marketing manager's historical and current involvement with trade shows may influence his/her performance evaluations.

Given the posterior probabilities of membership in the two derived clusters, i.e., the $\hat{p}_{ik}$'s, a logit transformation was performed on the probability that a marketing manager evaluated trade show performance primarily on a selling dimension. Specifically, the log was taken of the ratio of the selling cluster membership probability divided by one minus the selling cluster membership probability (adjustments of adding/subtracting a small positive constant were made for $\hat{p}_{ik} = 0$ or 1). Multiple regression analysis was performed using the 20 independent variables listed in Table 12. Because of missing data for some of the independent variables, 102 firms were included in this analysis. The results of this analysis are presented in Table 13. (Stepwise multiple regression analysis was also performed for more parsimonious results. The results were congruent with those in Table 13.) Five variables were significant in the equation: high technology products, new product introductions, sales concentration, importance to top management, and percent of promotion budget. Specifically, marketing managers who are most likely to emphasize selling results from trade show participation are with firms that sell high tech products with frequent new product introductions, have low customer sales concentrations, allocate a low percent of the sales promotion budget devoted to trade shows, and have top management who consider trade shows very important to the organization's success in meeting its marketing objectives.

An alternative approach for evaluating the practical usefulness of these 20 independent variables is to determine how well they can predict whether a marketing manager evaluates trade shows primarily on a selling or non-selling basis. Marketing managers were placed in either a selling or non-selling group based on which probability was higher. This procedure resulted in 48 managers being placed in the selling group and the remaining 54 managers categorized as non-selling (given the missing data). Two group multiple stepwise discriminant analysis was used to distinguish statistically between the two groups. The resulting discriminant function contained ten significant variables and produced a Wilks' lambda of .388 with chi-square of 63.312 ($p < .0001$). The ten variables in order of significance are (see Table 14) product technology, frequency of new product introduction, sales concentration, importance to top management, percent of promotion budget, selling information processing equipment, sales growth, marketing management's involvement, selling raw materials, and written trade show objectives. The results are quite similar to those presented in Table 13 concerning the logit regression analysis. These results indicate, in comparison with marketing managers who evaluate trade shows primarily on non-selling dimensions, those evaluating on selling dimensions: (1) sell more highly technical products, (2) frequently introduce new products, (3) do not concentrate their sales to a few large firms, (4) have their top management consider trade

TABLE 13

Logit Transformed Regression Results for Selling Evaluations

Variable	Beta	t
High tech products	32	2.794**
New product introduction	.34	2.862**
Sales concentration (%)	-.20	1.958*
Importance to top management	.23	1.920*
Percent of promotion budget	-.22	1.928*
Sales growth	.02	.169
Marketing manager's involvement	.05	.441
Written show objectives	.19	1.678
Industry life cycle	-.08	.656
Product modification (%)	-.12	.974
Marketing manager's experience	-.02	.177
Number of customers	-.17	1.345
Size	.05	.356
Selling raw materials	.15	1.377
Selling component parts	-.18	1.584
Selling major capital equipment	-.09	.742
Selling operating supplies	-.01	.106
Selling consumer durables	-.05	.448
Selling consumer nondurables	.07	.542
Selling information processing equipment	-.14	1.181
F = 2.217**		

*p < .05
**p < .01

shows to be more important for achieving the firm's marketing objectives, (5) spend less on trade shows as a percent of total promotion budget, (6) are not likely to be selling information processing equipment, (7) experience higher sales growth, (8) are firms in which the marketing manager is more intimately involved in trade show related decisions, (9) are more likely to be selling raw materials, and (10) are more likely to have written objectives for their overall trade show program.

TABLE 14

Discriminant Analysis of Performance Evaluation Groups

Variable	Mean Values: Selling Evaluation	Mean Values: Non-Selling Evaluation	Standardized Discriminant Function Coefficient[a]
High tech products	4.42	3.33	.7215
Frequent new product introduction	2.42	1.67	.6439
High sales concentration (%)	24.42	39.47	-.6426
Important to top management	3.92	3.31	.6329
Percent of promotion budget (%)	14.06	20.69	-.5844
Selling information processing equipment (%)	9.21	11.94	.4746
Sales growth (%)	20.95	14.47	.3309
Marketing manager's involvement	25.11	22.86	.3047
Selling of raw materials (%)	12.71	6.13	.2867
Written show objectives	1.27	1.58	.1880
Canonical Correlation	.781		
Wilks Lambda	0.388		
Chi square	63.312		
Significance	.001		
Correct classification rate (%)	81.37		

[a]All variables are significant at the .01 level or higher.

The above results concerning the profile of firms in which trade shows are primarily evaluated for their selling effectiveness appears to be consistent and logical. In general, selling oriented firms have a story to tell (e.g., new and high tech products), have a wide audience to reach, have written objectives because trade shows produce results that are quantifiable and central to the success of the organization, display intense marketing involvement, and have the support of top management. Discussion of these results with industry experts indicate that the industry results are also consistent in that manufacturers of information processing equipment, especially the larger organizations, do not actively sell on the trade show floor, while selling is quite common for marketers of raw materials. The most surprising result is that selling oriented organizations spend less as a percent of total sales promotion budget on trade show participation. This may reflect the cost efficiency of trade shows versus traditional field sales force selling (Mee 1982).

Here, 81.3% of the organizations were correctly classified as evaluating their trade show programs on a selling versus non-selling basis. This percentage of correct classifications was compared to a proportional chance criterion of .466 (Morrison 1969). Using a test of the difference between two proportions, $Z = 7.857$ which is extremely significant ($p \leq .001$). As a test of the upward bias in the classification results caused by reuse of the sample data, the Lachenbruch (1967, 1975) holdout procedure was used to classify individual organizations. The validated classification rate was 76.84%. This further indicates the predictor variables are important discriminators in this application.

6. Discussion

The conditional mixture maximum likelihood methodology for clusterwise linear regression has been technically described as well as the E-M algorithm for estimation. A Monte Carlo analysis investigating the performance of the methodology as a number of data and model factors were experimentally varied was presented. Finally, an application to trade show performance evaluations collected from senior marketing managers illustrated how two different groups of managers utilized very different criteria to evaluate their promotional expenditures in trade shows.

There are clearly other potential applications for this new methodology. For example, this clusterwise linear regression methodology could be utilized in the general context of multiattribute models for attitude measurement. In a similar vein, the E-M based procedure could be adapted for use in conjoint analysis studies (Green and Rao 1971) to investigate the basis of preference or choice. More substantive applications exist in virtually all the social sciences. In psychological testing, for example, this methodology could be utilized to identify groups of respondents that perform particularly poorly/well on specific items of a test. Concerning management research, the methodology could be utilized to relate firm strategy to resulting corporate performance and identify "strategic groups" (Porter 1980) or clusters of firms that utilize profiles of strategy to attain similar performance. Finally, in the area of political science, the procedure could be used to group countries with respect to common factors producing political risk levels (cf. Krayenbuehl 1985).

In addition, the methodology can be extended in a number of directions. For example, this conditional mixture approach could be modified to accommodate a binary choice or a rate dependent variable involving mixtures of other distributions from the exponential family. Like Basford and McLachlan (1985), the procedure can be generalized to accommodate three-way analyses where, for example, $\mathbf{y}$ and $\mathbf{X}$ could be given for various different

time periods or over various experimental manipulations. Another area of potential research involves modifying the procedure so user stipulated constraints as discussed in DeSarbo and Mahajan (1984) can be enforced (cf. DeSarbo, Oliver and Rangaswamy 1988). Finally an interesting generalization would be to accommodate the estimation of multiple (by cluster) ultrametric or path length tree(s) from such data.

References

ADDELMAN, S. (1962), "Orthogonal Main Effects Plans for Asymmetrical Factorial Experiments," *Technometrics, 4*, 21-46.

AKAIKE, H. (1974), "A New Look at Statistical Model Identification," *IEEE Transactions on Automatic Control, AC-19*, 716-723.

BASFORD, K.E., and MCLACHLAN, G.J. (1985), "The Mixture Method of Clustering Applied to Three-Way Data," *Journal of Classification, 2*, 109-125.

BINDER, D.A. (1978), "Bayesian Cluster Analysis," *Biometrika, 65*, 31-38.

BONOMA, T.V. (1983), "Get More Out of Your Trade Shows," *Harvard Business Review, 61*, 75-83.

BOZDOGAN, H. (1983), "Determining the Number of Component Clusters in Standard Multivariate Normal Mixture Models Using Model-Selection Criterion," *Technical Report VIC/DOM/A83-1*, Army Research Office.

CAVANAUGH, S. (1976), "Setting Objectives and Evaluating the Effectiveness of Trade Show Exhibits," *Journal of Marketing, 40*, 100-103.

CHARLIER, C.V.L., and WICKSELL, S.D. (1924), "On the Dissection of Frequency Functions," *Arkiv för Matematik, Astronomi Och Fysik., Bd. 18, No. 6*, 85-98.

CLEAVER, J. (1982), "You Don't Have to be a Star in this Show," *Advertising Age, 53*, 9.

COHEN, A.C. (1967), "Estimation in Mixtures of Two Normal Distributions," *Technometrics, 9*, 15-28.

COOPER, P. W. (1964), "Non Supervised Adaptive Signal Detection and Pattern Recognition," *Information and Control, 7*, 416-444.

DAY, N.E. (1969), "Estimating the Components of a Mixture of Normal Distributions," *Biometrika, 56*, 463-474.

DEMPSTER, A.P., LAIRD, N.M., and RUBIN, D.B. (1977), "Maximum likelihood from Incomplete Data Via the E-M Algorithm," *Journal of the Royal Statistical Society, B39*, 1-38.

DESARBO, W.S. (1982), "GENNCLUS: New Models for General Nonhierarchical Clustering Analysis," *Psychometrika, 47*, 449-476.

DESARBO, W.S., and CARROLL, J.D. (1985), "Three Way Metric Unfolding Via Alternating Weighted Least Squares," *Psychometrika, 50*, 275-300.

DESARBO, W.S., CARROLL, J.D., CLARK, L.A., and GREEN, P.E. (1984), "Synthesized Clustering: A Method for Amalgamating Alternative Clustering Bases with Differential Weighting of Variables," *Psychometrika, 49*, 57-78.

DESARBO, W.S., and MAHAJAN, V. (1984), "Constrained Classification: The Use of a Priori Information in Cluster Analysis," *Psychometrika, 49*, 187-215.

DESARBO, W.S., OLIVER, R., and RANGASWAMY, A. (1988), "A Simulated Annealing Methodology for Clusterwise Linear Regression," *Working Paper*, University of Michigan, Ann Arbor, MI.

DE SOETE, G., DESARBO, W.S., FURNAS, G.W., and CARROLL, J.D. (1984), "The Representation of Nonsymmetric Rectangular Proximity Data by Ultrametric and Path Length Tree Structure," *Psychometrika, 49*, 289-310.

De SOETE, G., DESARBO, W.S., and CARROLL, J.D. (1985), "Optimal Variable Weighting for Hierarchical Clustering: An Alternating Least Squares Algorithm," *Journal of Classification, 2*, 173-192.

DUNN, S.W., and BARBAN, A.M. (1986), *Advertising*, 6th ed., Hinsdale, IL: Dryden.

DYNKIN, E.B. (1961), "Necessary and Sufficient Statistics for a Family of Probability Distributions," *Selected Translations in Mathematical Statistics and Probability, Providence, RI: American Mathematical Society, 17-40.*

EVERITT, B.S., and HAND, D.J. (1981), Finite Mixture Distribution, New York: Chapman and Hall.

GANESALINGAM, S., and MCLACHLAN, G.J. (1981), "Some Efficiency Results for the Estimation of the Mixing Proportion in a Mixture of Two Normal Distributions," *Biometrics, 37*, 23-33.

GREEN, P.E., and RAO, V.R. (1971), "Conjoint Measurement for Quantifying Judgmental Data," *Journal of Marketing Research, 8*, 355-363.

GREEN, P.E., and TULL, D.S. (1978), *Research for Marketing Decisions*, 4th ed., Englewood Cliffs, NJ: Prentice-Hall.

GUILFORD, J.A. (1954), *Psychometric Methods*, 2nd ed., New York: McGraw-Hill.

HAAS, R.W. (1982), *Industrial Marketing Management*, 2nd ed., Boston: Kent.

HARTIGAN, J.A. (1975), *Clustering Algorithms*, New York: Wiley.

HARTIGAN, J.A. (1977), "Distribution Problems in Clustering," in *Classification and Clustering, ed. J. Van Ryzin, New York: Academic Press, 45-71.*

HASSELBLAD, V. (1966), "Estimation of Parameters for a Mixture of Normal Distributions," Technometrics, 8, 431-444.

HATCH, M. (1981), *How to Improve Sales Success at Trade Shows*, New Canaan, CT: Trade Show Bureau.

HOSMER, D.W. (1974), "Maximum Likelihood Estimates of the Parameters of a Mixture of Two Regression Lines," *Communications in Statistics, 3*, 995-1006.

HUTT, M.D., and SPEH, T.W. (1985), *Industrial Marketing Management*, 2nd ed., Hinsdale, IL: Dryden.

JOHNSTON, J. (1984), *Econometric Methods*, 3rd ed., New York: McGraw-Hill.

JUDGE, G.C., GRIFFITHS, W.E., HILL, R.C., LUTKEPOHL, H., and LEE, T.C. (1985), *Theory and Practice of Econometrics*, New York: Wiley.

KERIN, R.A., and CRON, W.L. (1987), "Assessing Trade Show Functions and Performance: An Exploratory Study," *Journal of Marketing, 51*, 87-94.

KONIKOW, R.B. (1983), *How to Participate Profitably in Trade Shows*, rev. ed., Chicago: Dartnell.

KRAYENBUEHL, T.E. (1985), *Country Risk: Assessment and Monitoring*, Lexington, MA: Lexington.

LACHENBRUCH, P.A. (1967), "An Almost Unbiased Method of Obtaining Confidence Intervals for the Probability of Misclassification in Discriminant Analysis," *Biometrics, 23*, 639-645.

LACHENBRUCH, P.A. (1975), *Discriminant Analysis*, New York: Hafner Press.

LILIEN, G.L. (1983), "A Descriptive Model of the Trade Show Budgeting Decision Process," *Industrial Marketing Management, 12*, 25-29.

LOUIS, T.A. (1982), "Finding the Observed Information Matrix When Using the E-M Algorithm," *Journal of the Royal Statistical Society, B44*, 226-233.

MACQUEEN, J. (1967), "Some Methods for Classification and Analysis of Multivariate Observations," in the *5th Berkeley Symposium of Mathematics, Statistics and Probability*, Vol. 1, eds. L.M. LeCam and J. Neyman, Los Angeles, CA: University of California Press, 281-298.

MADDALA, G.S. (1976), *Econometrics*, New York: McGraw-Hill.

MARRIOTT, F.H.C. (1975), "Separating Mixtures of Normal Distributions," *Biometrics, 31*, 767-769.

MCLACHLAN, G.J. (1982), "The Classification and Mixture Maximum Likelihood Approaches to Cluster Analysis," in *Handbook of Statistics*, Vol. 2, eds. P.R. Krishnaiah and L.N. Kanal, Amsterdam: North-Holland, 199-208.

MCLACHLAN, G.J. (1987), "On Bootstrapping the Likelihood Ratio Test Statistic for the Number of Components in a Normal Mixture," *Applied Statistics, 36*, 318-324.

MCLACHLAN, G.J. and BASFORD, K.E. (1988), *Mixture Models: Inference and Applications to Clustering*, New York: Marcel Dekker.

MEE, W.W. (1982), *Trade Show Exhibit Cost Analysis*, East Orleans, MA: Trade Show Bureau.

MEE, W.W. (1983a), *Trade Show Industry Growth 1972-1981, Projected Growth 1981-1991*, East Orleans, MA: Trade Show Bureau.

MEE, W.W. (1983b), *The Exhibitors: Their Trade Show Practices*, East Orleans, MA: Trade Show Bureau.

MEE, W.W. (1984), *Audience Characteristics — Regional and National Trade Shows*, East Orleans, MA: Trade Show Bureau.

MORRISON, D. (1969), "On the Interpretation of Discriminant Analysis," *Journal of Marketing Research, 6*, 156-163.

PEARSON, K. (1894), "Contribution to the Mathematical Theory of Evolution," *Philosophical Transactions of the Royal Society of London, A, 185*, 71-110.

PETERS, B.C., and WALKER, H.F. (1978), "An Iterative Procedure for Obtaining Maximum Likelihood Estimates of the Parameters for a Mixture of Normal Distributions," *SIAM Journal on Applied Mathematics, 35*, 362-378.

PORTER, M.E. (1980), *Competitive Strategy*, New York: Free Press.

QUANDT, R.E. (1972), "A New Approach to Estimating Switching Regressions," *Journal of the American Statistical Association, 67*, 306-310.

QUANDT, R.E., and RAMSEY, J.B. (1978), "Estimating Mixtures of Normal Distributions and Switching Regressions," *Journal of the American Statistical Association, 73*, 730-738.

REDNER, R.A., and WALKER, H.F. (1984), "Mixture Densities, Maximum Likelihood, and the E-M Algorithm," *SIAM Review, 2*, 195-239.

RICH, M. (1985), "Regional Shows Give Small Marketers an Even Break," *Marketing News*, May 27, 19, p. 15.

SCLOVE, S.C. (1977), "Population Mixture Models and Clustering Algorithms," *Communication in Statistics, A6*, 417-434.

SCLOVE, S.L. (1983), "Application of the Conditional Population Mixture Model to Image Segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-5*, 428-433.

SCLOVE, S.L. (1987), "Application of Model-Selection Criteria to Some Problems in Multivariate Analysis," *Psychometrika, 52*, 333-343.

SCOTT, A.J., and SYMONS, M.J. (1971), "Clustering Methods Based on Likelihood Ratio Criteria," *Biometrics, 27*, 238-397.

SPÄTH, H. (1979), "Algorithm 39: Clusterwise Linear Regression," *Computing, 22*, 367-373.

SPÄTH, H. (1981), "Correction to Algorithm 39: Clusterwise Linear Regression," *Computing, 26,* 275.

SPÄTH, H. (1982), "Algorithm 48: A Fast Algorithm for Clusterwise Linear Regression," *Computing, 29,* 175-181.

SPÄTH, H. (1985), *Cluster Dissection and Analysis,* New York: Wiley.

SYMONS, M.J. (1981), "Clustering Criteria and Multivariate Normal Mixtures," *Biometrics, 37,* 35-43.

TEICHER, H. (1961), "Identifiability of Mixtures," *Annals of Mathematical Statistics, 32,* 244-248.

TEICHER, H. (1963), "Identifiability of Finite Mixtures," *Annals of Mathematical Statistics, 34,* 1265-1269.

THEIL, H. (1971), *Principles of Econometrics,* New York: Wiley.

TITTERINGTON, D.M., SMITH, A.F.M., and MAKOV, U.E. (1985), *Statistical Analysis of Finite Mixture Distributions,* New York: Wiley.

VEAUX, R.D. (1986), "Parameter Estimation for a Mixture of Linear Regressions," *Technical Report No. 247,* Department of Statistics, Stanford University, Stanford, CA.

WILSON, D.L., and SARGENT, R.G. (1979), "Some Results of Monte Carlo Experiments in Estimating the Parameters of the Finite Mixed Exponential Distribution," in the *Proceedings of the Twelfth Annual Symposium Interface, ed. J.F. Gentleman, University of Waterloo, Ontario, Canada, 461-465.*

WOLFE, J.H. (1965), "A Computer Program for the Maximum Likelihood Analysis of Types." Technical Bulletin, 65-15, U.S. Naval Personnel Research Activity, San Diego, CA.

WOLFE, J.H. (1967), "NORMIX: Computational Methods for Estimating the Parameters of Multivariate Normal Mixtures of Distributions," *Research Memorandum SRM 68-2,* U.S. Naval Personnel Research Activity, San Diego, CA.

WOLFE, J.H. (1970), "Pattern Clustering by Multivariate Mixture Analysis," *Multivariate Behavioral Research, 5,* 329-350.

WOLFE, J.H. (1971), "A Monte Carlo Study of the Sampling Distribution of the Likelihood Ratio for Mixture of Multinormal Distributions," *Technical Bulletin: Naval Personnel and Training Research Laboratory, STB 72-2,* San Diego, CA.

YAKOWITZ, S.J. (1970), "Unsupervised Learning and the Identification of Finite Mixtures," *IEEE Transactions on Information Theory and Control, IT-16,* 330-338.

YAKOWITZ, S.J., and SPRAGINS, J.D. (1968), "On the Identifiability of Finite Mixtures," *Annals of Mathematical Statistics, 39,* 209-214.